EXHIBIT D

BOIES SCHILLER FLEXNER LLP
David Boies (*pro hac vice*)
333 Main Street
Armonk, NY 10504
(914) 749-8200
dboies@bsfllp.com

Maxwell V. Pritt (SBN 253155)
Joshua I. Schiller (SBN 330653)
Joshua M. Stein (SBN 298856)
44 Montgomery Street, 41st Floor
San Francisco, CA 94104
(415) 293-6800
mpritt@bsfllp.com
jischiller@bsfllp.com
jstein@bsfllp.com

Jesse Panuccio (*pro hac vice*)
1401 New York Ave, NW
Washington, DC 20005
(202) 237-2727
jpanuccio@bsfllp.com

David L. Simons (*pro hac vice*)
55 Hudson Yards, 20th Floor
New York, NY 10001
(914) 749-8200
dsimons@bsfllp.com

JOSEPH SAVERI LAW FIRM, LLP
Joseph R. Saveri (SBN 130064)
Cadio Zirpoli (SBN 179108)
Christopher K.L. Young (SBN 318371)
Holden Benon (SBN 325847)
Aaron Cera (SBN 351163)
Margaux Poueymirou (SBN 356000)
601 California Street, Suite 1505
San Francisco, California 94108
(415) 500-6800
jsaveri@saverilawfirm.com
czirpoli@saverilawfirm.com
cyoung@saverilawfirm.com
hbenon@saverilawfirm.com
acera@saverilawfirm.com

Matthew Butterick (SBN 250953)
1920 Hillhurst Avenue, #406
Los Angeles, CA 90027
(323) 968-2632
mb@buttericklaw.com

LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
Rachel Geman (*pro hac vice*)
250 Hudson Street, 8th Fl.
New York, NY 10013
(212) 355-9500
rgeman@lchb.com

Counsel for Individual and Representative Plaintiffs and the Proposed Class.

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

RICHARD KADREY, et al.,

Individual and Representative Plaintiffs,

v.

META PLATFORMS, INC.,

Defendant.

CASE NO. 3:23-cv-03417-VC

DECLARATION OF DR. JONATHAN L. KREIN IN SUPPORT OF PLAINTIFFS' OMNIBUS BRIEFING RE: EXISTING WRITTEN DISCOVERY

I, Dr. Jonathan L. Krein, declare as follows:

1. I have personal knowledge of the matters stated herein and, if called upon, can competently testify thereto. I make this declaration pursuant to 28 U.S.C. § 1746 and Local Rule 6-3 in support of Plaintiffs' Omnibus Briefing re: Existing Written Discovery.

2. I am a data scientist and software engineer with broad expertise in machine learning and language models, artificial intelligence, and software engineering. I was retained by Plaintiffs in this case to provide expert testimony primarily predicated on my inspection of Meta's source code associated with its Llama Large Language Models ("Llama Models") in relation to data collection and processing, model training and fine tuning, prompt engineering, and the creation of any scripts or other components used to prevent the models from regurgitating copyrighted material.

3. The work I have conducted on this case between inception and approximately October 4, 2024, was set forth in my October 4, 2024, Declaration in Support of Plaintiffs' Reply to Motion to Amend the Case Management Schedule. *See* Dkt. No. 206 (attaching my CV as Ex. A). In that Declaration, at a high level, I explained that, based on a review of the three (3) Source Code Repositories, I was missing materials such as "pull requests reflecting interactions between developers and source code repositories, commits reflecting modifications to source code repositories (*i.e.*, code changes), and additional source code repositories not yet produced." I gave specific examples of these missing materials, but naturally could not be certain that my examples encompassed everything that was missing. My examples included the following:

- With certainty, "source code related to how Meta trains its models to identify copyrighted material in order to prevent its models from regurgitating that material (or otherwise accomplishes the same, such as through tool use)."
- Possibly, "[redacted]
- Possibly: "[redacted]
- With certainty, "the production model code … including the application code that would encapsulate and/or exercise the production model code … [including] APIs

> that facilitate access between the user interface and the production llama model(s)… There must also be user interface code, among many other components."

4. I also explained in my October 4 Declaration that "Meta [had just] produced additional source code, pull requests, and commits on October 2, 2024, that [would] require additional time to review and digest." I have since reviewed that additional material, but find that important components, including the items described in my October 4 Declaration and listed above, are still missing. Additionally, since October 4, no new material has been made available on the source code review machines (the Source Code Computers).

5. In light of those facts, I have undertaken additional efforts to identify and describe what is missing. It is important to recognize that AI systems are not like traditional software applications, and it is significantly more difficult to discern what is missing from the development record of an AI system than it is for a traditional application.[1] I have worked to identify the missing material since first receiving access to the source code. Early on it was clear that things were missing, but it takes time to study the materials available in order to infer more precisely what exactly is missing. To be clear, there is a finite set of materials, which Meta should have in its possession, and with certainty that full set of materials has not been produced.

6. Therefore, I remain of the view that at least some, if not all of the materials listed in my October 4 Declaration plainly exist. Obviously, not having access to Meta's files, I cannot be fully certain of everything that is missing, nor of where it is all located (file names, *etc*.). For instance, if there are other materials in a similar category, but that were not referred to in the materials to which I have access, I would not know about them. Thus, production of additional

[1] AI systems are challenging because their development entails numerous piecemeal components that do not directly reference on another. Development of an AI system includes experimental source code, research source code, model training source code, and model evaluation source code, among other things. Each of these broad areas of development entail numerous pods of source code that do not reference each other. And this is all in addition to the AI system itself, which is a traditional software application (or collection of interworking applications). Note that a "model" is simply a collection of numbers that correspond to a mathematical structure and a set of algorithms. Conversely, an AI system, such as Meta.ai, houses and runs a model as part of its execution, but encompasses more than just a core model.

materials may reveal yet other materials that are missing. Since I can only point to what I can so far see, the only entity in a position to ensure full production in an efficient manner is Meta.

7. In addition to the materials I described in my October 4 Declaration and restated in the preceding paragraphs, the ongoing review of extant Repositories has revealed concrete references (*e.g.*, URLs and file paths) to additional important materials that have not been made available to me. To be clear, these are not new materials I am raising for the first time here in this declaration. As I explained in my October 4 declaration, "I believe that Meta likely has not made available to Plaintiffs all relevant source code associated with its Llama Models—including … additional source code repositories not yet produced," among other things. The "new" materials are listed in Appendix A.

8. In addition to the materials listed in Appendix A and discussed above and in my October 4 Declaration, files containing [redacted] for Llama models 2, 3, and 4 also appear to be missing. The source code that has been produced includes a [redacted] which Meta created for the purpose [redacted] [redacted].[2] The source code contains results of Meta's [redacted] analysis for Llama 1, but the results for Llama 2, 3, and 4 are missing.[3]

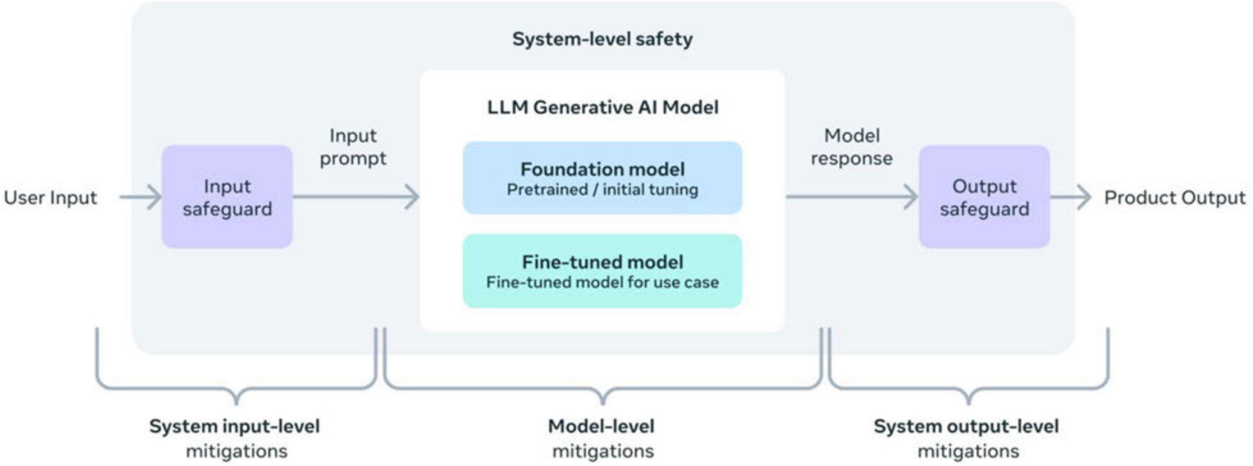
9. Finally, as explained in my October 4 Declaration, the application software for the running AI system (including all of its infrastructure components) is important and still has not been provided. As is well known in the industry, an AI system includes more than just an AI model. The following graphic from Meta's own documentation shows an example of this.[4] Although this particular graphic focuses on safety—and thus does not show all of the components that exist in such systems—it shows both that additional components exist and that such components are important to an understanding of such systems in the context of copyright. For example, none of the source code produced by Meta includes the "Input safeguard" and "Output safeguard"

[2] *See* the directory [redacted]

See [redacted]

[4] *See* https://ai.meta.com/static-resource/responsible-use-guide/, p. 5.

components from this diagram, despite that Meta included this diagram in its documentation as a representation of its Llama models. Meta's source code production also lacks any code for the user interface of Meta.ai, or any of the tools the AI interfaces with, among other things, as part of providing the services found at https://www.meta.ai/. These additional components, and the system as a whole, are important to understanding, at the very least, issues of memorization, as they determine what happens between when the model generates output to when—and what—output is finally shown to the user.



10. As I explained in my October 4 Declaration, the "application [*i.e.*, system and infrastructure] source code is relevant for at least three reasons: (i) it is the code defining the system that customers actually use; (ii) it is organized into a connected and complete flow, whereas the science and engineering code is, as typically expected, found in numerous disjoint pieces; and (iii) it likely contains relevant components that may not be found in the science and engineering code."

I declare under penalty of perjury that the foregoing is true and correct. Executed on the 6th day of November, 2024, in San Francisco, California.

By: /s/ *Dr. Jonathan L. Krein*
Dr. Jonathan L. Krein

APPENDIX A

Additional missing materials include at least the following.[5]

- **Source code repositories** that exist in [redacted], including at the same or similar location as the three repositories so far produced. Note that, to the extent these repositories have been excluded from production, but clearly exist, others, including located in the same [redacted] also exist, as I asserted in my October 4 Declaration. Additionally, these repositories would be private to the account, and so would not be visible without proper credentials.
    - [redacted] there are at least 22 references to this repository.[6]
    - [redacted] there is at least one reference to this repository.[7]
    - [redacted] – there are at least three references to this repository.[8]
    - [redacted]– there are at least three references to this repository.[9]
    - [redacted] – there are at least two references to this repository.[10]
    - [redacted] There are at least twelve references to this repository.[11]
- **Data directories** that are part of the source code, being utilized by the source code to perform the various steps of the model construction process, as the source code clearly shows. These data directories are referring to specific data files other than those produced so far in this case.
    - [redacted] there are at least 887 references to this data directory, which includes, but is not limited to:
        - [redacted] – there are at least three references to this data directory, including, for instance:

[5] References to source code that begin with [redacted]” are from the latest commit of the source code found in the *second* source code production. References that lack [redacted] are to the latest [redacted] found in the *first* production.

[6] *See, e.g.*, [redacted]

- [redacted].[12]
- [redacted].[13]
  - [redacted]– there are at least 430 references to this data directory.[14]
- [redacted]– there are at least 2,271 references to this data directory, which includes, but is not limited to:
  - [redacted] there are at least 673 references.[15]
  - [redacted]– there are at least 1,075 references.[16]
  - [redacted] – there are at least 605 references.[17]
- [redacted]– there are at least 430 references.[18]

- **Data files** used in the training/finetuning of models to make them safer, such as to prevent models from outputting offensive language. Of particular importance are the data files pertaining to [redacted]. These include at least the following files (some paths are partial or missing where unavailable in the source code produced; one or more of these files may overlap with the data directories mentioned above):
  - [redacted]
  - [redacted][20]

[12] [redacted]

○ [redacted]

[redacted] [redacted]

[redacted] [redacted]

[redacted] [redacted]

[redacted] [redacted]

[redacted] [redacted]

[redacted] [redacted]

[redacted] [redacted]

[redacted]